Тематический раздел: Теоретическая и компьютерная химия. **Полная исследовательская публикация**
Подраздел: Физическая органическая химия. *Регистрационный код публикации:* 11-27-14-1
Публикация доступна для обсуждения в рамках функционирования постоянно действующей интернет-конференции "*Бутлеровские чтения*". http://butlerov.com/readings/
Поступила в редакцию 21 мая 2011 г. УДК 544.354.081.7:004.021.

Методы хемоинформатики в термодинамике равновесий. Диссоциация уксусной кислоты.

© **Бондарев Сергей Николаевич**,[1] **Зайцева Инна Сергеевна**[2]
и **Бондарев Николай Васильевич**[1]*+
[1] *Харьковский национальный университет им. В.Н. Каразина. пл. Свободы, 4. г. Харьков, 61077. Украина. E-mail: n_bondarev@ukr.net*
[2] *Харьковская национальная академия городского хозяйства. Ул. Маршала Бажанова, 17. г. Харьков, 61002. Украина. E-mail: inna.zaitseva123@yandex.ru*

*Ведущий направление; +Поддерживающий переписку

Ключевые слова: *хемоинформатика, искусственная нейронная сеть, множественная регрессия, массив данных, константа диссоциации, энергия Гиббса, сольватация.*

Аннотация

Методы химической информатики (хемоинформатики) – множественная линейная регрессия и нейросетевое моделирование применены для анализа зависимости энергии Гиббса (констант) диссоциации уксусной кислоты от свойств водно-органических растворителей. Выявлены значимые факторы, влияющие на равновесие диссоциации кислоты. Построена нейросетевая модель (трехслойный персептрон) и показана перспективность применения нейронных сетей для прогнозирования констант диссоциации (силы) уксусной кислоты в водно-органических растворителях.

Введение

В физикохимии растворов, как и в химии в целом, накоплено огромное количество экспериментальных данных, проведение глубокого анализа которых уже невозможно без применения средств современной информатики – "науки о принципиально новой человеко-машинной технологии расширенного воспроизводства качественно нового знания" [1].

Вследствие этого, на стыке химии и информатики возникает и быстро оформляется в самостоятельную научную дисциплину *хемоинформатика* [2-5], методы которой начинают активно внедряться во все области химии. Терминология дисциплины еще не устоялась: хемоинформатика (chemoinformatics), хеминформатика (cheminformatics, chemiinformatics), химическая информатика (chemical informatics) [3].

Узкое, но очень распространенное понимание хемоинформатики – применение методов информатики в биоорганической химии для создания лекарств [2]. В дальнейшем эта дефиниция была расширена. В частности, согласно определению, данному Г. Пэризом (Paris, 2000), хемоинформатика – это научная дисциплина, охватывающая дизайн, создание, организацию, управление, поиск, анализ, распространение, визуализацию и использование химической информации [3], в предмет исследования которой включены приемы хранения, извлечения и обработки химической информации.

Развитию хемоинформатики в значительной мере способствует наличие обоснованной методологии и реализующего ее программного обеспечения, которые позволяют химику на основе обработки экспериментальных данных осуществлять прогнозирование самых разнообразных свойств химических соединений и процессов [4-7]. При этом на первый план выходят методы нелинейного моделирования (в частности, нейросетевые технологии прогнозирования [8]) как многообещающие в области анализа массивов данных и прогнозирования свойств сложных систем.

В отличие от статистических моделей, которые построены на том, что вначале делается предположение о характере связей между анализируемыми переменными, а затем проверяется соответствие данных предложенной модели, при нейросетевом моделировании (neural networks) не делается предположений об истинной форме этих связей.

Новизна данной работы сводится к демонстрации согласованности традиционных подходов (регрессионно-корреляционный, сольватационно-термодинамический) с современными нейросетевыми методами анализа и прогнозирования термодинамических свойств химических равновесий. Причем сольватационно-термодинамический метод [9] дает возможность выявить вклады эффектов среды (сольватации реагентов) в изменение термодинамических параметров химических равновесий.

Особенностью регрессионно-корреляционного подхода является принцип линейности свободных энергий (ЛСЭ), позволяющий выявить причины (электростатические, химические взаимодействия) влияния растворителя на термодинамику химических реакций [10]. Область применения нейронных сетей во многом совпадает с кругом задач, решаемых традиционными статистическими методами.

Однако, по сравнению с *линейными методами* статистики (линейная регрессия, авторегрессия, линейный дискриминант), нейросети позволяют эффективно строить *нелинейные зависимости*.

Искусственные нейронные сети (ИНС) широко применяются в органической, аналитической, физической и биологической областях химии для обработки массивов экспериментальных данных с целью построения прогностических моделей [11-13].

Специфика нейросетевого подхода к анализу и прогнозированию данных. Преимущество ИНС перед классическими методами статистического анализа заключается в возможности аппроксимации по экспериментальным данным любых сколь угодно сложных *нелинейных* зависимостей произвольного и заранее неизвестного вида [14].

Другая существенная особенность нейронных сетей состоит в том, что зависимость между входным и выходными данными находится в процессе обучения сети [15]. Кратко рассмотрим основные понятия теории нейронных сетей применительно к многослойному персептрону (МП) [8].

Понятие нейрона. ИНС состоят из определенного количества «искусственных нейронов» [16]. Нейрон имеет несколько каналов ввода информации – дендриты и канал вывода информации – аксон. Аксон нейрона соединен с дендритами других нейронов с помощью синапсов.

На рис. 1 представлена графическая модель нейрона, из которого видно, что через несколько входных каналов j-й нейрон получает сигналы $x(i)$ от других нейронов, каждый из которых умножается на $w(j,i)$ – вес синаптической связи выхода нейрона i с входом нейрона j, положительное значения которого соответствуют *возбуждающим* синапсам, отрицательные значения – *тормозящим* синапсам; если $w(j,i) = 0$, то связь между нейронами j и i отсутствует. Далее производится сложение преобразованных сигналов (блок сумматор СУМ) и добавляется порог возбуждения (активации) $b(j)$ j-го нейрона.

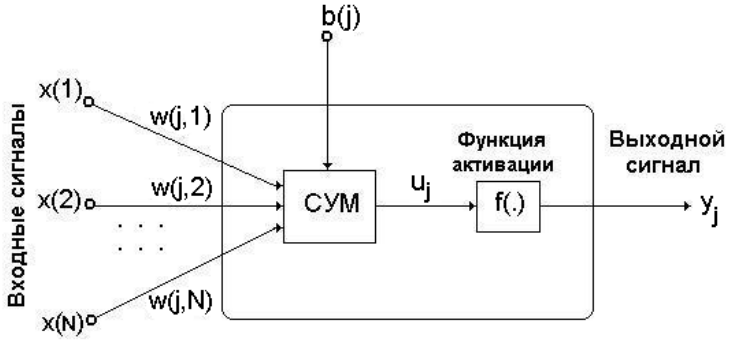


Рис. 1. Схема искусственного нейрона

Текущее состояние нейрона (индуцированное локальное поле нейрона j) описывается соотношением:

$$u_j = \sum_{i=1}^{N} w(j,i)x(i) + b(j) \qquad (1)$$

где $x(i)$ – входные сигналы, $i = 1, 2, \ldots, N$. Индекс j относится к номеру рассматриваемого нейрона в сети, индекс i указывает номер синаптической связи.

Полученный нейроном сигнал преобразуется с помощью *нелинейной* функции активации или передаточной функции $f(.)$ в *выходной* сигнал

$$y_j = f(u_j) \qquad (2)$$

Входной слой нейронов служит для ввода значений входных переменных, выходной слой – для вывода результатов. Кроме того, в сети может быть еще много промежуточных (скрытых) нейронов, выполняющих внутренние функции. Последовательность слоев нейронов (входных, скрытых и выходных) и их соединений называется *архитектурой* сети.

Любой нейрон скрытого или выходного слоя персептрона может выполнять два типа вычислений [8]:

- вычисление сигнала на выходе нейрона, реализуемое в виде непрерывной нелинейной функции от входного сигнала и весов синаптических связей данного нейрона;
- вычисление градиента поверхности ошибки весов синаптических связей данного нейрона, необходимого для обратного прохода через сеть.

Функция активации нейронов *f*(.) – это некоторая нелинейная функция, моделирующая процесс передачи возбуждения. Самой популярной формой функции является сигмоидальная. Наличие нелинейности функции необходимое условие, так как в противном случае отражение "вход-выход" сети можно свести к обычному однослойному персептрону [17].

Важнейшее свойство многослойного персептрона заключается в способности обучаться аппроксимации любых сколь угодно сложных нелинейных зависимостей между входными и выходными данными [18-20]. Алгоритмы обучения позволяют найти веса синаптических связей и пороговые значения активации нейронов, с помощью которых минимизируется ошибка прогноза сети [21].

Начальная конфигурация (архитектура) сети выбирается случайным образом, и процесс обучения прекращается либо когда пройдено определенное количество эпох (эпохе в итерационном обучении нейронной сети соответствует один проход по всему обучающему множеству с последующей проверкой на контрольном множестве), либо когда ошибка достигает некоторого определенного уровня малости или вообще перестает уменьшаться.

Целью работы является применение линейного многофакторного и нейросетевого подхода к анализу взаимосвязи между энергиями Гиббса диссоциации уксусной кислоты и свойствами водно-органических растворителей для построения прогностических зависимостей и нейросетевых моделей.

Экспериментальная часть

Анализируемые данные: зависимые переменные – аквамоляльные [22] стандартные энергии Гиббса ($\Delta G^{o}_{d,HAc}$) диссоциации уксусной кислоты [23-25]; независимые переменные – физико-химические свойства водно-органических растворителей (вода – метанол, вода – этанол, вода – пропан-2-ол) – диэлектрическая проницаемость, электроноакцепторные параметры Димрота-Райхардта, электронодонорные параметром Камлета-Тафта и плотность энергии когезии [26-29].

Множественный линейный анализ. Для установления взаимосвязи между энергией Гиббса диссоциации уксусной кислоты ($\Delta G^{o}_{d,HAc}$) и свойствами водно-органических растворителей – нормализованными параметрами Димрота-Райхардта(E_T^N) и Камлета-Тафта (B_{KT}), диэлектрической проницаемостью (ε^N) и плотностью энергии когезии (δ^2_N) на основе линейной модели первого порядка:

$$\Delta G^{o}_{d,HAc} = b_0 + b_1(1/\varepsilon^N) + b_2 E_T^N + b_3 B_{KT} + b_4 \delta^2_N \qquad (3)$$

Для проведения анализа использованы данные по энергиям Гиббса диссоциации уксусной кислоты в растворителях вода–метанол, вода–этанол, вода–пропан-2-ол при 298.15 К с содержанием неводного компонента до 0.7 мол. доли (область пересольватации [30], после которой изменяется состав сольватных оболочек ионов и молекул кислоты и проявляется нелинейность зависимости $\Delta G^{o}_{d,HAc}$ от свойств растворителей).

Результаты и их обсуждение

Полученные значимые дескрипторы и наличие мультиколлинеарности выявлялись с помощью корреляционной матрицы (табл. 1) и пошаговым включением параметров в уравнение регрессии. Критерием включения дескриптора в уравнение регрессии был выбран уровень значимости ($p < 0.05$). Коэффициенты уравнения множественной регрессии оценены стандартным методом наименьших квадратов (МНК).

Статистически значимыми факторами, определяющими зависимость энергии Гиббса диссоциации уксусной кислоты от свойств растворителей, являются диэлектрическая проницаемость и плотность энергии когезии (табл. 1).

Табл. 1. Полные коэффициенты парной корреляции

N = 22	$1/\varepsilon_N$	E_T^N	B_{KT}	δ^2_N	$\Delta_r G^o_d$(HAc)
$1/\varepsilon_N$	1.0000	-0.8744	0.8778	-0.6597	0.9335
E_T^N	-0.8744	1.0000	-0.9774	0.5789	-0.8438
B_{KT}	0.8778	-0.9774	1.0000	-0.7082	0.9008
δ^2_N	-0.6597	0.5789	-0.7082	1.0000	-0.8635
$\Delta_r G^o_d$(HAc)	0.9335	-0.8438	0.9008	-0.8635	1.0000

Коэффициент корреляции между независимыми переменными $1/\varepsilon_N$ и δ^2_N составляет –0.6597, то есть эти дескрипторы слабо коррелированны, поэтому дальнейший регрессионный анализ проведен с учетом только этих дескрипторов.

В табл. 2 приведены результаты анализа для уравнения регрессии:

$$\Delta G^o_d(\text{HAc}) = (37.3 \pm 2.8) + (3.62 \pm 0.51)\cdot \varepsilon_N^{-1} - (13.3 \pm 2.7)\cdot \delta^2_N \qquad (4)$$

Для проверки статистического качества оцененного уравнения регрессии использована стандартная процедура, детально описанная в работах [31, 32]: проверка адекватности уравнения регрессии экспериментальным данным по критерию Фишера; проверка статистической значимости коэффициентов уравнения регрессии по критерию Стьюдента; проверка свойств данных, выполнимость которых предполагалась при оценивании двухпараметрического уравнения: математическое ожидание случайного отклонения равно нулю для всех наблюдений, постоянство дисперсии отклонений, отсутствие автокорреляции остатков (критерий Дарбина-Уотсона), отсутствие мультиколлинеарности (матрица полных коэффициентов корреляции), ошибки имеют нормальное распределение (детальный анализ остатков), отсутствие выбросов (расстояния Махалонобиса и показатель Кука).

Как следует из значений множественного коэффициента корреляции и детерминации (табл. 2), построенная модель с двумя факторами $1/\varepsilon_N$ и δ^2_N адекватно описывает экспериментальные данные $\Delta_r G^o_d$(HAc).

Табл. 2. Итоги регрессии для зависимой переменной $\Delta_r G^o_d$(HAc)

R = 0.99 R^2 = 0.98 F(2,19) = 454.02 Стандартная ошибка оценки: 0.64						
	Бета коэффициенты	Стандартная ошибка	B	Стандартная ошибка	t(19)	p-уровень
			37.3	1.4	27.2	0.00
$1/\varepsilon_N$	0.64	0.04	3.6	0.2	14.7	0.00
δ^2_N	-0.44	0.04	-13.3	1.3	-10.0	0.00

99% (R^2 = 0.98) дисперсии зависимой переменной $\Delta_r G^o_d$(HAc) объясняется влиянием независимых дескрипторов. Критерий Фишера F имеет высокое значение $F(2.19) = 454.02$ против $F_{кр}(2.19) = 3.52$, что подтверждает статистическую значимость линейной модели регрессии.

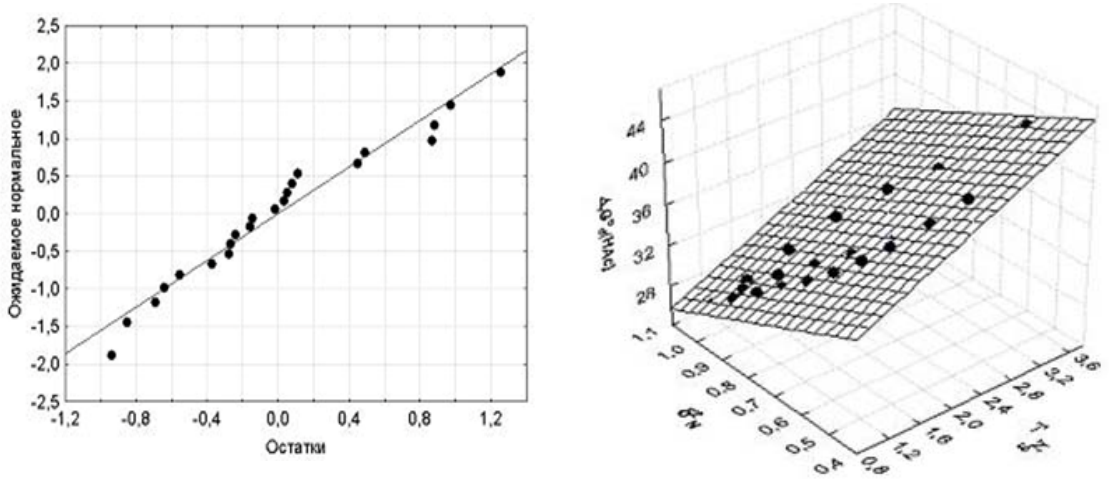
Применение полных коэффициентов парной корреляции при множественной регрессии (табл. 1) для изучения связи двух величин может привести к неправильным выводам. Поэтому проанализированы частные коэффициенты корреляции, отражающие степень линейной взаимосвязи между двумя переменными, вычисленную после устранения влияния всех других факторов. Частные коэффициенты корреляции приведены в табл. 3.

Табл. 3. Частные коэффициенты корреляции и толерантность

	Независимая переменная $\Delta_r G^o_d$(HAc)						
	Бета коэффициенты	Частная корреляция	Получастная корреляция	Толерантность	R-квадрат	t(19)	p-уровень
$1/\varepsilon_N$	0.64	0.96	0.48	0.56	0.44	14.7	0.00
δ^2_N	-0.44	-0.92	-0.33	0.56	0.44	-10.0	0.00

Как видно из табл. 1, 3, при пошаговом исключении переменных $1/\varepsilon_N$ и δ^2_N коэффициент корреляции между $\Delta_r G^o_d$(HAc) и независимыми переменными повышается: полный и частный коэффициенты корреляции соответственно равны 0.93 и 0.96; -0.86 и -0.92.

Это означает, что оба предиктора незначительно маскируют истинную взаимосвязь исследуемых переменных. Толерантность для предикторов $1/\varepsilon_N$ и δ^2_N составляет (табл. 3) 0.56, что свидетельствует об отсутствии мультиколлинеарности (избыточности предикторов).



а) б)

Рис. 2. Нормальный вероятностный график остатков (а) и график зависимости энергии Гиббса диссоциации уксусной кислоты $\Delta_r G^o_d$(HAc) от свойств водно-органических растворителей ε^{-1}_N и δ^2_N (б)

Табл. 4. Сопоставление наблюдаемых значений энергии Гиббса диссоциации уксусной кислоты с предсказанными по уравнению (4)

мол. доля DO или DMSO	Свойства растворителя				$\Delta G^o_{d,HAc}$ наблюдаемые	$\Delta G^o_{d,HAc}$ предсказанные	Остатки
	$1/\varepsilon^N$	E_T^N	B_{KT}	δ^2_N			
Вода – диоксан							
0	1	1	0.19	1	27.15	27.62	-0.47
0.1	1.645	0.77	0.36	0.917	32.56	31.06	1.50
0.2	2.625	0.69	0.45	0.834	38.06	35.71	2.35
0.3	3.848	0.62	0.48	0.751	44.02	41.24	2.78
0.4	5.326	0.60	0.50	0.668	50.83	47.70	3.13
0.5	9.229	0.54	0.51	0.584	58.88	62.94	-4.06
Вода – диметилсульфоксид							
0	1	1	0.19	1	27.15	27.62	-0.47
0.1	1.025	0.85	0.35	0.931	29.51	28.63	0.88
0.2	1.047	0.75	0.45	0.862	33.03	29.63	3.40

На рис. 2 приведен нормальный вероятностный график остатков (рис. 2а) и трехмерный график зависимости энергии Гиббса диссоциации уксусной кислоты $\Delta_r G^o_d$(HAc) от свойств водно-органических растворителей $\varepsilon^{-1}{}_N$ и $\delta^2{}_N$ (рис. 2б), свидетельствующие об адекватности выбранной линейной модели: остатки удовлетворяют нормальному распределению, так как они располагаются вблизи линии нормального распределения (а); практически все точки лежат на плоскости (б).

При использовании уравнения (2) для оценки энергии Гиббса диссоциации уксусной кислоты в водно-диоксановых и водно-диметилсульфоксидных растворителях выявлена удовлетворительная прогностическая способность уравнения (2) для водно-органических растворителей с малым содержанием диоксана либо диметилсульфоксида табл. 4.

Нейросетевое оценивание. Данные для нейронной сети: свойства водно-органических растворителей (входных переменные), энергии Гиббса диссоциации уксусной кислоты (выходных переменные).

Используя иллюстративные возможности интерактивной компьютерной графики [1], построены (рис. 3) поверхности зависимостей энергии Гиббса диссоциации уксусной кислоты от свойств водно-органических растворителей вода–метанол, вода–этанол и вода–пропан-2-ол) с шагом 0.1 мол. доля неводного компонента. Сложный вид поверхностей (рис. 3) свидетельствует в пользу проведения нейросетевого моделирования, так как факт нелинейности задачи не вызывает сомнения. Конечно, можно было бы попробовать решить задачу статистическими методами нелинейного анализа. Однако, как уже отмечено ранее, осложняющим обстоятельством является необходимость формулировки гипотезы о явном виде изучаемой зависимости, которая, как это видно из рис. 3, совсем не является очевидной.

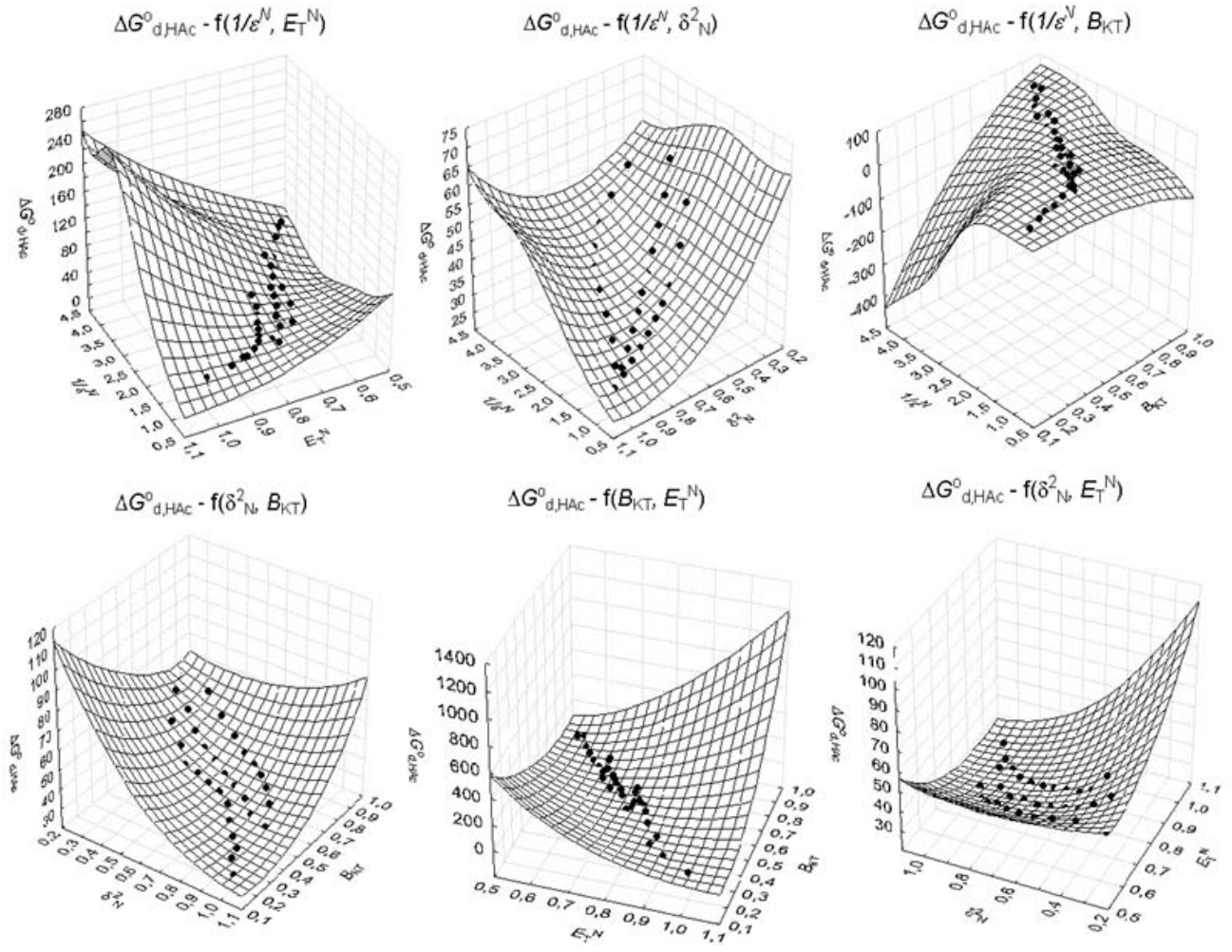


Рис. 3. Поверхности $\Delta G^o_{d,HAc} = f(1/\varepsilon^N, E_T{}^N)$, $\Delta G^o_{d,HAc} = f(1/\varepsilon^N, \delta^2{}_N)$, $\Delta G^o_{d,HAc} = f(1/\varepsilon^N, B_{KT})$, $\Delta G^o_{d,HAc} = f(\delta^2{}_N, B_{KT})$, $\Delta G^o_{d,HAc} = f(B_{KT}, E_T{}^N)$, $\Delta G^o_{d,HAc} = f(\delta^2{}_N, E_T{}^N)$.

Так как известны входные и выходные вектора сети, то для ее обучения можно использовать алгоритмы обучения с учителем [8, 21]. При этом за счет способности к обобщению сетью могут быть получены новые результаты, если подать на вход данные (свойства водно-органических растворителей), которые не использовались при обучении сети.

С помощью процедуры *понижение размерности* выявлено, что все свойства водно-органических растворителей существенно влияют на силу уксусной кислоты (энергию Гиббса диссоциации).

Выбор структуры (архитектуры) нейронной сети (сколько использовать промежуточных слоев и нейронов в них) является наиболее сложной проблемой. Известны различные методики для выбора оптимальной структуры сети [8, 21] однако, в большинстве случаев, их применимость к той или иной задаче сильно зависит от качества и количества входных данных.

Предварительно были проанализированы 1000 сетей (линейная, радиальная базисная функция (число скрытых элементов *min* = 1, *max* = 8); трехслойный персептрон (число скрытых элементов *min* = 1, *max* = 10) и выбран перспективный тип сети и вариант архитектуры – это трехслойный персептрон с пятью скрытыми нейронами МП 4:4-5-1:1 (табл. 5, рис. 4).

В рассматриваемой регрессионной задаче линейная нейросетевая модель характеризуется самой низкой производительностью (табл. 5) значение стандартного отклонения на контрольной выборке равно 0.424, для модели радиальной базисной функции (РБФ) это значение составляет 0.222.

Табл. 5. Статистические характеристики нейросетей разного типа

Архитектура, коэффициент корреляции Пирсона	Производительность обучения	Контр. производительность	Тест. производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Обучение/ Элементы*
Линейная 2:2-1:1, 0.8923	0.417	0.424	0.465	0.139	0.082	0.198	ПО
РБФ 4:4-7-1:1 0.9873	0.120	0.253	0.222	0.013	0.010	0.029	КС,КБ,ПО
МП 4:4-5-1:1, 0.9947	0.090	0.114	0.119	0.030	0.014	0.048	ОР100, СГ20, СГ31b
МП 4:4-8-1:1, 0.9965	0.056	0.107	0.117	0.018	0.013	0.049	ОР100, СГ20, СГ52b

* *Примечание.* Алгоритмы (коды), использованные для оптимизации сетей: код ПО – псевдообратные (линейная оптимизация методом наименьших квадратов); код КС – К-средних (расстановка центров); код КБ – К-ближайших соседей (задание отклонений); код ОР – обратное распространение; код СГ – метод сопряженных градиентов; b – код остановки (сеть с наименьшей ошибкой на контрольной выборке) [33]. Код СГ31b показывает, что для оптимизации сети использован метод сопряженных градиентов и что сеть найдена на 31 эпохе по минимальной ошибке на валидационном множестве.

Сеть с архитектурой МП 4:4-8-1:1 имеет несколько лучшие статистические характеристики (контрольная производительность равна 0.107) по сравнению с МП 4:4-5-1:1 (контрольная производительность равна 0.114), однако в случае небольших массивов данных предпочтение отдается менее сложным архитектурам сетей.

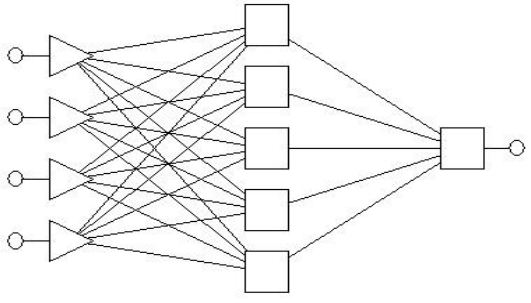


Рис. 4. Архитектура трехслойного персептрона с прямой передачей сигнала для прогнозирования констант диссоциации (энергии Гиббса диссоциации) уксусной кислоты по свойствам водно-органических растворителей

Следует заметить, что построенные таким образом сети не обязательно лучшие из всех возможных, так как при использовании нелинейных возможностей нейромоделирования с целью минимизации ошибки никогда нет уверенности в том, что нельзя добиться еще меньшей ошибки [21]. В особенности это относится к многослойным персептронам.

Поэтому дальнейшая корректировка сети была проведена с помощью алгоритмов быстрого распространения (100 эпох на первом этапе) и Левенберга-Маркара (500 эпох на втором этапе), который считается одним из лучших алгоритмов нелинейной оптимизации [33]. Алгоритм быстрого распространения является эвристической модификацией алгоритма обратного распространения, где для ускорения сходимости применяется простая квадратичная модель поверхности ошибок, вычисляемых для каждого веса синаптической связи. Алгоритм Левенберга-Маркара применяется только для относительно небольших сетей с одним выходом, что в наибольшей степени соответствует условиям нашей задачи.

В работе также были проанализированы персептроны МП 4:4-3-1:1 и МП 4:4-4-1:1, содержащие в скрытом слое соответственно три и четыре нейрона (табл. 6). Несмотря на достаточно высокую способность к аппроксимации данных, эти сети обладают меньшей способностью к обобщению (прогнозированию).

Табл. 6. Результаты обучения сетей МП 4:4-3-1:1 и МП 4:4-4-1:1

Архитектура	Производительность обучения	Контрольная производительность	Тестовая производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Обучение/ Элементы
МП 4:4-3-1:1	0.054	0.073	0.074	0.018	0.024	0.017	ОР100,СГ20,СГ55b
МП 4:4-4-1:1	0.093	0.092	0.160	0.028	0.020	0.036	ОР100,СГ20,СГ29b

Для обучения сетей все множество наблюдений было разбито на три выборки (по умолчанию осуществлялось случайное разделение наблюдений между выборками) во избежание переобучения сети и для гарантирования качественного обобщения (прогнозирования). Первая из них (*Обучающая* – 50% наблюдений) использовалась для обучения сети; вторая (*Контрольная* – 25% наблюдений) – для кросс-валидации алгоритма обучения во время его работы, и третья (*Тестовая* – 25% наблюдений) – для окончательного независимого тестирования. Обучение проведено со скоростью 0.01.

В качестве активационной функции на промежуточных слоях при нейросетевом моделировании использована функция гиперболического тангенса (tanh), сигмоидальная нелинейность которой определяется следующим образом [8]:

$$\tanh(u) = \frac{\exp(au) - \exp(-au)}{\exp(au) + \exp(-au)}, \text{ где } a > 0. \qquad (5)$$

На первой эпохе алгоритм быстрого распространения корректирует веса синаптических связей по формуле обобщенного дельта-правила [34], как и в методе обратного распространения [8]:

$$\Delta w_{j,i}(n) = \eta \delta_j o_i + \alpha \Delta w_{j,i}(n-1) \qquad (6)$$

где n – номер примера обучения; η – скорость обучения, используемая при переходе от одного шага процесса к другому (была выбрана равной 0.01); δ_j – локальный градиент ошибки; α – как правило, положительное значение, называемое постоянной момента или коэффициентом инерции (был выбран равным 0.3); o_i – выходное значение i-го нейрона.

На последующих эпохах алгоритм использует предположение о квадратичности поверхности ошибок для более быстрого продвижения к точке минимума. Изменения весов вычисляются по формуле быстрого распространения [33]:

$$\Delta w_{j,i}(n) = \frac{y_j(n)}{y_j(n-1) - y_j(n)} \Delta w_{j,i}(n-1) \qquad (7)$$

где $y_j(n)$ – выдаваемое сетью выходное значение, соответствующее n-му примеру обучения.

Коррекция весов методом Левенберга-Маркара проводится по формуле [33]:

$$\Delta w_{j,i}(n) = -\left(Z^T Z + \lambda I\right)^{-1} Z^T \boldsymbol{\varepsilon} \qquad (8)$$

где $\boldsymbol{\varepsilon}$ – вектор ошибок на всех наблюдениях; Z – матрица-якобиан, содержащая первые частные производные ошибок нейронной сети по переменным весов и смещений весов синаптических связей; λ – параметр алгоритма, определяемый в процессе линейной (скалярной) оптимизации вдоль выбранного направления.

Первый член в формуле Левенберга-Маркара соответствует линейной модели, а второй – градиентному спуску. Управляющий параметр I задает относительную значимость этих двух вкладов.

Физико-химическая интерпретация параметров множественной линейной регрессии. В полученной линейной двухфакторной модели (см. уравн. 2), характеризующем зависимость энергии Гиббса диссоциации кислоты ΔG^{o}_{d}(HAc) от обратной величины диэлектрической проницаемости ε_N^{-1} и плотности энергии когезии δ^2_N, параметр $b_1 > 0$, а $b_2 < 0$. Следовательно, с уменьшением диэлектрической проницаемости и уменьшением плотности энергии когезии при замене воды на водно-органические растворители сила кислоты уменьшается.

Это означает, что:

- уменьшение диэлектрической проницаемости растворителя сопровождается уменьшением энергии Гиббса сольватации ионов, следствием чего является сдвиг равновесия в сторону недиссоциированных молекул кислоты;
- при уменьшении плотности энергии когезии водно-органические растворители в большей мере стабилизируют недиссоциированную форму уксусной кислоты (рис. 5).

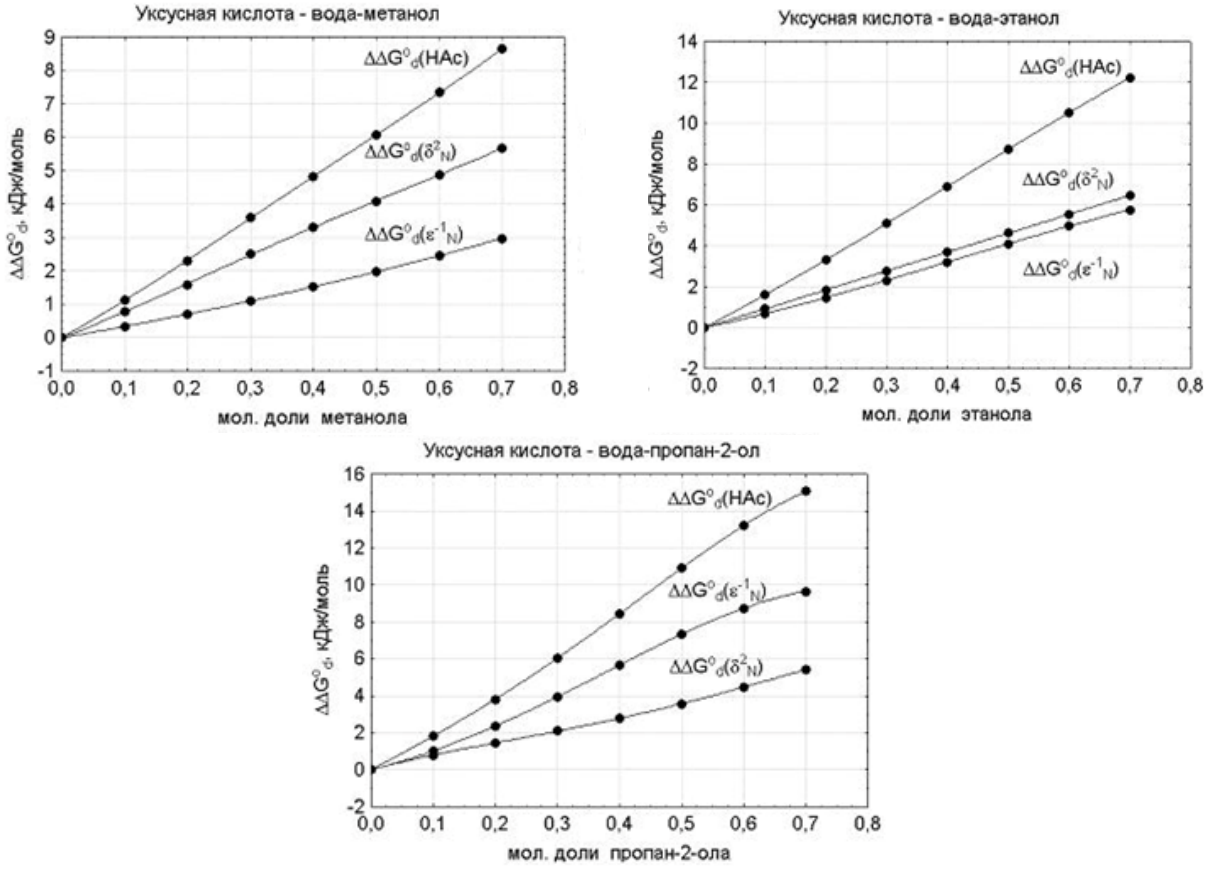


Рис. 5. Влияние плотности энергии когезии и диэлектрических свойств смешанных растворителей вода–метанол, вода–этанол и вода–пропан-2-ол (w-s) на уменьшение силы уксусной кислоты ($\Delta\Delta G^{o}_{d}$(HAc) = $\Delta G_d^{o,w\text{-}s}$(HAc) - $\Delta G_d^{o,w}$(HAc))

Сопоставление результатов линейного многофакторного регрессионного анализа (рис. 5) и результатов сольватационно-термодинамического анализа (рис. 6 в качестве примера [35])

позволяют говорить об их адекватности. Уменьшение силы уксусной кислоты в водно-метанольных растворителях по сравнению с водой обусловлено возрастанием сольватации недиссоциированных молекул кислоты и уменьшением сольватации ацетат-ионов (рис. 6).

После оптимизации обучения многослойного персептрона МП 4:4-5-1:1 (с 4 входными, 5 скрытыми и 1 выходным нейронами, рис. 4) с помощью алгоритмов быстрого распространения и Левенберга-Маркара (табл. 7), сеть имеет лучшие статистические показатели, чем приведенные в табл. 6. Так, контрольная производительность сети увеличилась с 0.886 (1-0.114) до 0.960 (1-0.040), ошибка на валидационной выборке уменьшилась с 0.013 до 0.011, коэффициент корреляции Пирсона возрос с 0.9947 до 0.9984.

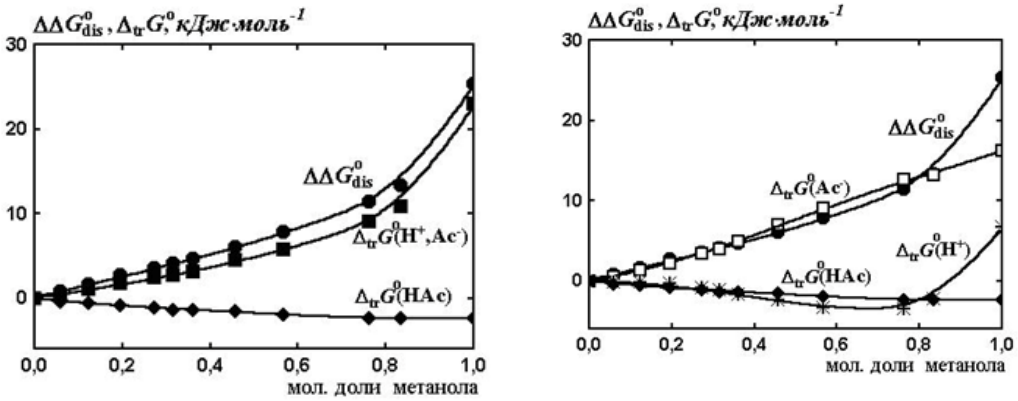


Рис. 6. Влияние энергии Гиббса пересольватации ионов ($\Delta_{tr}G^o(H^+)$, $\Delta_{tr}G^o(Ac^-)$ и молекул $\Delta_{tr}G^o(HAc)$ на изменение энергии Гиббса диссоциации уксусной кислоты $\Delta\Delta G^o_d(HAc)$ при замене воды на водно-метанольные растворители

Табл. 7. Результаты обучения персептрона МП 4:4-5-1:1

Архитектура, коэффициент корреляции Пирсона	Производительность обучения	Контрольная производительность	Тестовая производительность	Ошибка обучения	Контрольная ошибка	Тестовая ошибка	Обучение/ Элементы
МП 4:4-5-1:1 0.9984	0.059	0.040	0.084	0.017	0.011	0.017	БР100,ЛМ235b

Опишем необходимые технические детали, связанные с обучением нейросети на использованных в работе данных.

Производительность обучения; *контрольная производительность*; *тестовая производительность* – это отношение стандартного отклонения ошибки прогноза к стандартному отклонению исходных данных на соответствующих выборках.

Ошибка обучения; *контрольная ошибка*; *тестовая ошибка* – ошибка сети на выборках, используемых во время обучения.

Ошибка для конкретной конфигурации сети (мера эффективности обучения) определяется путем прогона через сеть всех имеющихся наблюдений и сравнением реально выдаваемых выходных значений (y_i) с желаемыми (целевыми) значениями (d_i).

Сигнал ошибки $e_j(n)$ выходного нейрона j на итерации n, соответствующей n-му примеру обучения, рассчитывается по формуле:

$$e_j(n) = d_j(n) - y_j(n) \qquad (9)$$

В пакетном режиме обучения по методу обратного распространения среднеквадратическая ошибка сети определяется по уравнению [8]:

$$E_{av}(n) = \frac{1}{2N}\sum_{n=1}^{N}\sum_{j\in C} e_j^2(n), \qquad (10)$$

где множество C включает все нейроны выходного слоя сети, N – общее число образов (примеров) в обучающем множестве. Внутреннее суммирование по j выполняется по всем нейронам выходного слоя сети, в то время как внешнее суммирование выполняется по всем примерам данной эпохи.

В пакетном режиме корректировка веса $w_{j,i}$ осуществляется только после прохождения сети всего множества примеров. При последовательном режиме обучения по методу обратного распространения корректировка весов проводится после подачи каждого примера.

Отметим, что ошибка на валидационной выборке не превышает ошибку на независимой (тестовой) выборке, то есть выполняется необходимое условие оптимизации [21].

Как следует из табл. 6, результаты работы нейронной сети на трех множествах практически совпадают, что говорит о приемлемом качестве нейронной сети.

Обучение. На первом этапе использован алгоритм быстрого распространения (100 эпох). На втором этапе – метод Левенберга-Маркара. Сеть выбрана на 235 эпохе по минимальной ошибке на контрольной выборке (код остановки b).

Критериями качества обученной нейросетевой модели служат статистические характеристики, представленные в табл. 8.

Точки на графике расположены достаточно близко к прямой, лежащей под углом 45 градусов к осям координат, что свидетельствует об эффективной работе нейронной сети.

Иллюстрацией качества работы нейронной сети также является график зависимости наблюдаемых значений энергии Гиббса диссоциации уксусной кислоты (выходной переменной) $\Delta G^{o}_{d,HAc}$ ($\Delta G^{o}_{d,HAc}$-Observed) от предсказанных значений ($\Delta G^{o}_{d,HAc}$ -Predicted) (рис. 7) и табл. 9, 10.

Данные табл. 9 отражают качество работы нейронной сети, в обучение которой были включены энергии Гиббса диссоциации уксусной кислоты в водно-пропан-2-ольных растворителях, в то время как результаты табл. 10 являются прогнозом силы кислоты уже обученной модели по свойствам водно-диметилсульфоксидных растворителей.

На основании результатов табл. 4, 9, 10 можно сделать вывод о согласованности методов статистического и нейросетевого анализа термодинамики химических равновесий.

Следует отметить, что при прогнозировании констант диссоциации в водно-диметилсульфоксидных растворителях с большим содержанием диметилсульфоксида и в водно-диоксановых растворителях не удалось получить удовлетворительных результатов с помощью построенной в работе нейросети, что вполне объяснимо, так как сеть была обучена только на водно-спиртовых растворителях.

Табл. 8. Статистические показатели обученной нейросети для моделирования зависимости силы уксусной кислоты от свойств водно-органических растворителей*

	Обучающая выборка $\Delta G^{o}_{d,HAc}$	Контрольная выборка $\Delta G^{o}_{d,HAc}$	Тестовая выборка $\Delta G^{o}_{d,HAc}$	Общая выборка $\Delta G^{o}_{d,HAc}$
Среднее данных	38.29	40.86	33.30	37.71
Стандартное отклонение данных	8.69	7.94	4.72	8.18
Среднее ошибки	0.009	0.053	-0.301	-0.056
Стандартное отклонение ошибки	0.51	0.32	0.40	0.47
Среднее абсолютной ошибки	0.36	0.27	0.44	0.36
Отношение стандартных отклонений	0.059	0.040	0.084	0.057
Корреляция	0.9983	0.9992	0.9987	0.9984

* *Примечание: Среднее данных* – среднее значение исходных энергий Гиббса диссоциации уксусной кислоты $\Delta G^{o}_{d,HAc}$. В нашем случае они лежат в интервале 33-40, что свидетельствует о корректном разбиении массива данных на обучающую, контрольную и тестовую выборки. *Стандартное отклонение* – среднеквадратическое отклонение исходных данных $\Delta G^{o}_{d,HAc}$. *Среднее ошибки* – среднее значение ошибки прогноза (ошибка – это разница между исходными и рассчитанными $\Delta G^{o}_{d,HAc}$). *Стандартное отклонение ошибки* – стандартное отклонение ошибки прогноза $\Delta G^{o}_{d,HAc}$. *Среднее абсолютной ошибки* – средняя абсолютная ошибка прогноза (абсолютная ошибка – это разница (по модулю) между исходными и рассчитанными $\Delta G^{o}_{d,HAc}$. *Отношение стандартных отклонений* или производительность – отношение стандартного отклонения ошибки прогноза к стандартному отклонению исходных данных (Стандартное отклонение ошибки/Стандартное отклонение данных). *Корреляция* – коэффициент корреляции Пирсона.

Представляется перспективным применения нейронных сетей для анализа взаимосвязи термодинамики химических равновесий и физико-химических свойств водно-органических

растворителей (без ограничения состава смешанного растворителя) с целью прогнозирования констант равновесий (силы слабого электролита или устойчивости комплексов) в растворителях, для которых эти данные (константы равновесий) отсутствуют или их определение сопряжено с экспериментальными трудностями (например, в чистых диметилсульфоксиде или диоксане).

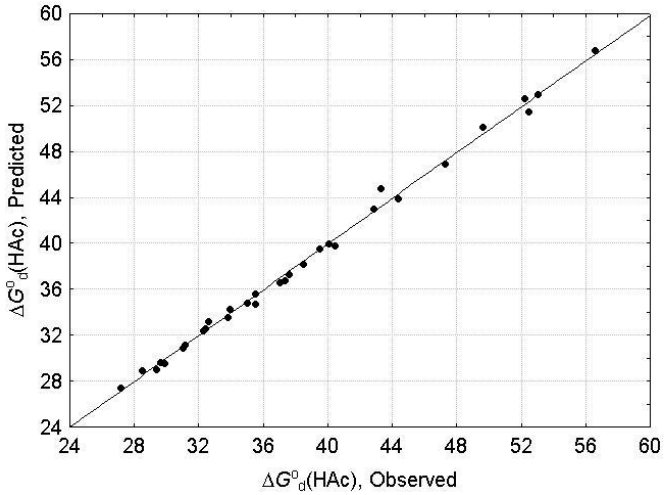


Рис. 7. Зависимость предсказанных нейросетью значений энергии Гиббса диссоциации уксусной кислоты $\Delta G^{o}_{d,}$(HAc), предсказанные (Predicted) от наблюдаемых $\Delta G^{o}_{d,}$(HAc), наблюдаемые (Observed) в водно-органических растворителях

Табл. 9. Иллюстрация качества работы нейросети на примере предсказания энергии Гиббса (кДж/моль) диссоциации уксусной кислоты по свойствам водно-пропан-2-ольных растворителей

мол. доля	Свойства растворителя				$\Delta G^{o}_{d,HAc}$	$\Delta G^{o}_{d,HAc}$	
пропан-2-ола	$1/\varepsilon^{N}$	E_T^{N}	B_{KT}	δ^2_N	наблюдаемые	предсказанные	Остатки
0	1.000	1.00	0.19	1.000	27.15	27.36	0.21
0.1	1.287	0.77	0.49	0.943	29.62	29.60	-0.02
0.2	1.650	0.70	0.60	0.889	32.40	32.54	0.14
0.3	2.087	0.67	0.65	0.839	35.02	34.80	-0.22
0.4	2.568	0.66	0.69	0.787	37.04	36.59	-0.45
0.5	3.032	0.64	0.72	0.730	38.49	38.21	-0.28
0.6	3.411	0.62	0.75	0.664	40.10	39.96	-0.14
0.7	3.668	0.60	0.79	0.586	42.86	43.02	0.16
0.8	3.829	0.58	0.83	0.492	47.30	46.88	-0.42
0.9	3.973	0.56	0.89	0.379	52.23	52.59	0.36
1.0	4.117	0.54	0.88	0.242	53.02	52.93	-0.09

Табл. 10. Иллюстрация качества работы нейросети на примере прогнозирования энергии Гиббса диссоциации уксусной кислоты по свойствам водно-диметилсульфоксидных растворителей

мол. доля	Свойства растворителя				$\Delta G^{o}_{d,HAc}$	$\Delta G^{o}_{d,HAc}$	
DMSO	$1/e^{N}$	E_T^{N}	B_{KT}	δ^2_N	наблюдаемые	предсказанные	Остатки
0	1.000	1.00	0.19	1.000	27.15	27.36	0.21
0.1	1.025	0.85	0.35	0.931	29.51	30.40	0.89
0.2	1.047	0.75	0.45	0.862	33.03	32.56	-0.47
0.3	1.091	0.69	0.50	0.793	38.83	35.13	-3.70

Одним из путей решения этой проблемы может быть пополнение массива данных по константам диссоциации и свойствам смешанных растворителей, содержащих как протолитический так и апротонный органический компонент для обучения сети на более репрезентативных выборках и расширения ее прогностических возможностей.

Выводы

1. Сочетание термодинамических и традиционных статистических методов химической информатики позволяет выявить не только вклады первичных эффектов среды (энергии Гиббса пересольватации реагентов) в изменение силы электролита, но и дает информацию о причинах смещения химического равновесия под влиянием растворителя.

2. Установлена двухпараметрическая зависимость констант (энергии Гиббса) диссоциации уксусной кислоты от диэлектрической проницаемости и плотности энергии когезии водно-органических растворителей (содержание спирта до 0.7 мол. доли метанола, этанола или пропан-2-ола), свидетельствующая о том, что а) уменьшение диэлектрической проницаемости растворителя, сопровождающееся уменьшением энергии Гиббса сольватации ионов, сдвигает равновесие диссоциации в сторону недиссоциированных молекул кислоты; б) при уменьшении плотности энергии когезии водно-органические растворители в большей мере стабилизируют недиссоциированную форму уксусной кислоты. Эти выводы согласуются с результатами сольватационно-термодинамического анализа – уменьшение силы уксусной кислоты в водно-метанольных растворителях по сравнению с водой обусловлено возрастанием сольватации недиссоциированных молекул кислоты и уменьшением сольватации ацетат-ионов.
3. Проведен нейросетевой анализ зависимости энергии Гиббса диссоциации уксусной кислоты от физико-химических свойств водно-органических растворителей (вода–метанол, вода–этанол, вода–пропан-2-ол).
4. Построена нейросетевая модель (трехслойный персептрон) и спрогнозированы константы диссоциации уксусной кислоты в водно-диметилсульфоксидных растворителях с содержанием органического компонента до 0.3 мол. доли.
5. Показана перспективность применения нейронных сетей для анализа взаимосвязи термодинамики химических равновесий и физико-химических свойств водно-органических растворителей с целью прогнозирования констант равновесий (силы слабого электролита или устойчивости комплексов) в растворителях, для которых эти данные отсутствуют.

Литература

[1] Зенкин А.А. Когнитивная компьютерная графика. *М.: Наука.* **1991**. 192с.
[2] F.K. Brown. Chapter 35. Chemoinformatics: What is it and How does it Impact Drug Discovery. *Annual Reports in Medicinal Chemistry.* **1998**. Vol.33. P.375-384.
[3] R. Leach Andrew, J. Gillet Valerie. An Introduction to Chemoinformatics. *Springer.* **2007**. 256p.
[4] B.A. Bunin, A. Siesel, G.A. Morales, J. Bajorath. Chemoinformatics: Theory, Practice, & Products. *Springer.* **2007**. 295p.
[5] I.I. Baskin, A. Varnek. Chapter 1. Fragment Descriptors in SAR/QSAR/QSPR Studies, Molecular Similarity Analysis and in Virtual Screening. In: Chemoinformatics Approaches to Virtual Screening. *Ed. RCS Publishing.* **2008**. P.1-43.
[6] Коняев Д.С. Методы анализа данных и химической информатики в исследовании комплексообразования в растворах и на поверхности химически модифицированных кремнеземов. *Дисс. … канд. хим. наук.* **1999**. С.139-141.
[7] Киреева Н.В. Прогнозирование констант устойчивости комплексов лантаноидов и щелочноземельных металлов с органическими лигандами и температур плавления ионных жидкостей методами химической информатики. *Автореф. дис. … канд. хим. наук. М.: ИФХЭ им. А.Н. Фрумкина РАН и ун-т Страсбурга (Франция).* **2010**. 25с.
[8] Хайкин С. Нейронные сети: полный курс. 2-е изд. *М.: Издательский дом "Вильямс".* **2006**. 1104с.
[9] Бондарев Н.В. Сольватационно-термодинамические эффекты водно-метанольного растворителя в координации катионов Na^+, K^+, NH_4^+ и Ag^+ с 18-краун-6. Энергии Гиббса комплексообразования и пересольватации реагентов. *ЖОХ*. **2006**. Т.76. Вып.1. С.13-18.
[10] Смирнова Е.В., Цыба Ю.В., Бондарев Н.В., Зайцева И.С. Регрессионный анализ влияния свойств водно-диметилсульфоксидных растворителей на силу уксусной и бензойной кислот. Materiały V Międzynarodowej naukowe-praktycznej konferencji "Naukowa przestrzeń Europy - 2009". *Przemyśl, Polska: Nauka i studia.* **2009**. Vol.7 (Chemia i chemiczne technologie). S.6-8.
[11] Баскин И.И. , Палюлин В.А., Зефиров Н.С. Применение искусственных нейронных сетей в химических и биохимических исследованиях. *Вестн. Моск. ун-та. Сер. 2. Химия.* **1999**. Т.40. №5. С.323-326.
[12] Гальберштам Н.М., Баскин И.И., Палюлин В.А., Зефиров Н.С. Нейронные сети как метод поиска зависимостей структура–свойство органических соединений. *Успехи химии.* **2003**. Т.72. №7. С.706-727.

[13] Краснянчин Я.Н., Пантелеймонов А.В., Холин Ю.В. Надежность идентификации аналитов с помощью искусственных нейронных сетей. *Вестн. Харьк. нац. ун-та.* **2010**. №895. Химия. Вип. 18(41). С.39-45.
[14] Баскин И.И. Моделирование свойств химических соединений с использованием искусственных нейронных сетей и фрагментных дескрипторов. *Автореф. дис…. докт. физ-мат. наук. М.: МГУ им. М.В. Ломоносова.* **2010**. 49с.
[15] Ежов А.А., Шумский С.А. Нейрокомпьютинг и его применения в экономике и бизнесе. *М.: МИФИ.* **1998**. 224с.
[16] Уоссермен Ф. Нейрокомпьютерная техника: Теория и практика. *М.: Мир*. **1992**. 118с.
[17] Розенблатт Ф. Принципы нейродинамики: персептроны и теория механизмов мозга. *М.: Мир*. **1965**. 480с.
[18] Галушкин А.И. Теория нейронных сетей. Т.1. Нейрокомпьютеры и их применение. *М.: ИПРЖР*. **2000**. 416с.
[19] Горбань А.Н., Дунин-Барковский В.Л., Кирдин А.Н.и др. Нейроинформатика. *Новосибирск: Наука. Сибирское предприятие РАН.* **1998**. 296с.
[20] Аксенов С.В., Новосельцев В.Б. Организация и использование нейронных сетей (методы и технологии). *Томск: Изд-во НТЛ.* **2006**. 128с.
[21] Боровиков В.П. Нейронные сети. *Statistica Neural Networks*. Методология и технологии современного анализа данных. 2-е изд. *М.: Горячая линия – Телеком*. **2008**. 392с.
[22] E.N. Tsurko, N.V. Bondarev. Mathematical modeling of solvent parameters' relevant contribution to the alpha-amino acid (valine, alpha-alanine) solvation in H_2O–MeOH, H_2O–EtOH and H_2O– PrOH-2. *J. Mol. Liquids.* **2007**. No.131-132. P.151-157.
[23] Никольский Б.П. Справочник химика. *Л.: Изд-во «Химия»*. **1965**. Т.3. Изд.2. 1008с.
[24] Лебедь В.И., Бондарев Н.В. Константы диссоциации и термодинамические характеристики диссоциации уксусной и бензойной кислот в смесях вода – метанол, вода – диоксан. *Журн. физ. химии.* **1982**. Т.56. №1. С.30-33.
[25] Лебедь В.И., Бондарев Н.В., Пауленова А. Константы диссоциации и термодинамические характеристики диссоциации и сольватации уксусной кислоты в смесях вода – пропанол-2. *Журн. физ. химии.* **1987**. Т.61. №6. С.1487-1491.
[26] Зайцева И.С., Ельцов С.В., Кабакова Е.Н., Бондарев Н.В. Корреляционный анализ влияния эффектов среды на энергетику комплексообразования катионов натрия и калия с эфиром 18-краун-6 в водно-органических растворителях. *Журн. общ. химии.* **2003**. Т.73. Вып.7. С.1079-1084.
[27] Афанасьев В.Н., Ефремова Л.С., Волкова Т.В. Физико-химические свойства бинарных растворителей. Водосодержащие системы. *Иваново: ИХР РАН*. **1988**. 412с.
[28] C. Kalidas, G. Hefter, Y. Marcus. Gibbs energies of transfer of cations from water to mixed aqueous organic solvents. *Chem. Rev.* **2000**. Vol.100. No.3. P.819-852.
[29] G. Hefter, Y. Marcus, W.E. Waghorne. Enthalpies of Transfer of Electrolytes and Ions between Water and Mixed Aqueous Solvents. *Chem. Rev.* **2002**. Vol.102. No.8. P.2773-2836.
[30] Гордон Д. Органическая химия растворов электролитов. *М.: Мир*. **1979**. 712с.
[31] Ларина О.В., Бондарев Н.В., Керн А.П. Эффекты среды и комплексообразование солей натрия, калия, аммония и серебра (I) с 18-краун-6 эфиром в водно–пропан-2-ольных растворителях. Химия. *Вестник Харьк. нац. университета.* **2007**. Вып.15(38). №770. С. 301-312.
[32] Цыба Ю.В., Бондарев Н.В., Зайцева И.С. Статистический анализ экспериментальных данных и термодинамика диссоциации карбоновых кислот. Materiály IV Mezinárodní vědecko-praktická konference "Věda: teorie a praxe - 2008". Praha, Czech Republic: Publishing House "Education and Science" s.r.o. **2008**. Dil 10 (Chemie a chemická technologie). P.12-16.
[33] StatSoft, Inc. Электронный учебник по статистике. *Москва, StatSoft*. **2001**. WEB: http://www.statsoft.ru/home/textbook/default.htm.
[34] B. Widrow, M.E. Hoff. Adaptive switching circuits. IRE WESTCON Conferencion Record. *New York.* **1960**. P.96-104.
[35] Цыба Ю.В., Бондарев Н.В., Зайцева И.С. Регрессионно-корреляционный и термодинамический анализ влияния свойств растворителей вода-метанол, вода-этанол, вода-пропан-2-ол на силу уксусной кислоты. Материали за V Международна научна практична конференция "Динамиката на съвременната наука - 2009". *София, България: "Бял ГРАД-БГ" ООД.* **2009**. Т.12. С.35-37.